The Impact of Gigabit Network Research on Scientific Visualization

Charles Hansen
Advanced Computing Laboratory

Stephen Tenbrink
Computer Networking

Los Alamos National Laboratory
Los Alamos, New Mexico 87545

Abstract

Networks based on the High Performance Parallel Interface (HIPPI) will become the norm at LANL. The ramifications of such a high speed networking paradigm on scientific visualization are enormous. Not only will scientist have the capability of networked framebuffer animation loops in their offices, but the partitioning of graphics tasks between MIMD, SIMD and specialized hardware will also be feasible. Of course, as bandwidth increases, the problem size quickly grows to exceed whatever the limits. For this reason, the investigation of gigabyte networks currently underway at Los Alamos National Laboratory.

Introduction

Los Alamos National Laboratory (LANL) has one of the world's largest supercomputer networks with several Cray XMPs, YMPs, TMC CM-2s and a CM-5 plus various other computer systems. Currently, these systems are networked based on a LANL designed 50 Mbit/s link called the High Speed Parallel Interface (HSPI). With the growing demand for imaging, visualization, data transfer and more efficient networks, Los Alamos is using its experience to build a new network based on gigabit/sec links [1].

This network is based on the ANSI standard High Performance Parallel Interface (HIPPI). This standard specifies a data rate of 800 and 1600 Mbits/s over a point to point simplex channel with a look ahead ready that eliminates the effect of longer transmission paths. Los Alamos will base it's Intergrated Computing Network (ICN) on multiple HIPPI channels networked through crossbar switches in concert with crossbar interfaces.

The ramifications of such a high speed networking paradigm on imaging and scientific visualization are enormous. Not only will scientist have the capability of networked framebuffer animation loops in their offices, but the partitioning of graphics tasks between MIMD, SIMD and specialized hardware will also be feasible.

Background

The Integrated Computer Network (ICN) at LANL, which connects the supercomputer and various other computer systems, was designed and built by LANL personal starting from the last part of the 1970's and continuing through the 1980s. The links between the

computers use a LANL designed interface called the High Speed Parallel Interface (HSPI) which operates at data rates up to 50 Mbit/s [2].

The current ICN is a store and forward packet switch network whose switches are large minicomputers. The network uses other minicomputers to act as concentrators of terminal traffic. Some of these concentrators contain specially designed hardware that supports data rates of up to 300 KBit/s to graphics terminals located in user's offices. These unique links offer the user full connectability to all network resources especially the Cray computers where they are used for visualization purposes. The data rate, however, limits the effectiveness of visualization due to the long latency times in imaging to the user's display. This latency is even more pronounced when the data transported via the network is images rather than geometry.

Another aspect of the current ICN is that it uses a LANL developed protocol called Simple Inter-Machine Protocol (SIMP). While this protocol has worked well in our terminal based network, it is incompatible with standard protocols that are vendor supplied with common workstations.

The present ICN architecture with its minicomputer switches and concentrators, coupled with SIMP, is proving to be a bottleneck when higher data rates are attempted to support visualization and imaging. When the data rate on the link between the concentrators and the user's terminal was increased to 500 Kbit/s no appreciable increase in aggregate throughput was noticed.

One solution that has worked for a small number of users is the use of frame buffers connected directly to the Cray channels via the HSPI paradigm [3]. This technique can support a 512 by 512 pixel display with 8 bits of color at 24 frames per second. While this frame buffer has proven quite popular with certain users, it is limited by the fact that only a very few can be supported by Cray's low speed channel. To support more than a few users a *network solution* is required.

HIPPI

The *network solution* for imaging has evolved to one that not just provides low resolution animations but also high resolutions of 1024 by 1024 pixel display with up to 24 bits of color at 30 frames per second. Such a network requires near gigabit/sec. links. The concept that Los Alamos came up with is an 800 Mbit/sec link coupled with appropriate switches.

Clearly, there are other reasons for the development of a gigabit computer network. Along with the imaging aspect, there is the general desire for network links to keep up with the increasing CPU performance. The current ICN with its 50 Mbit/sec links was efficient enought to support the CDC 6600's, 7600's and Cray 1's suring the late 70's and early 80's. However, with the current class of supercomputers, the ICN has seen bottlenecks occurring between these machines and our Common File System (CFS). In addition to file transport, there is a desire to run several nodes in parallel on certain applications (including MIMD and SIMD machines) by efficiently distributing processes on various supercomputers and possibly high performance workstations. To make this concept feasible, the basic network structure needs to provide the necessary foundation.

For the physical link, LANL found that , at that time (1987), there did not exist a standard for computer network channels in the near gigabit/sec arena. A preliminary proposal was presented to the ANSI X3T9 committee for such a link. The result of this effort is the High Performance Parallel Interface (HIPPI) channel that will soon become an official ANSI standard (X3.183-1991). At the same time, other standardization efforts have been developed for a data link layer and the networking layers for the HIPPI. In addition, techniques for transmitting the HIPPI data over fiber optics have evolved in Fiber Channel ANSI committee.

The HIPPI standard specifies a data rate of 800 and 1600 Mbits/sec over a point to point simplex channel with a look ahead ready that eliminates the effect of longer transmission paths. It also carries a pre-connection addressing capability (I-field) that allows it to be switched in a network environment. The networking protocols are currently being standardized and will use IEEE 802.2 control. This will allow the use of such upper layer protocols as TCP/IP to be implemented on HIPPI based networks.

Networking

It should be noted that HIPPI is a channel specification and does not specify or imply any set network architecture as does FDDI. It is possible to connect a number of HIPPI nodes in a ring or star network or just between two computers. The point to point nature eliminates multiple access like ethernet making the point to point connectivity more secure.

The point to point nature of HIPPI makes switches not only necessary but also very important. Using only multiple HIPPI connections directly to and from machines clearly isn't the best way to solve this. Using a serial switch in the network can limit the aggregate bandwidth through individual switches. Whereas at any point in time, a bus based switch, such as a minicomputer, can have only one transaction occurring on the bus, a crossbar switch can have simultaneous transactions. Los Alamos designed and developed a 16 by 16 HIPPI crossbar switch (CBS) in 1988-1989 that could support an aggregate data rate of 12.8 Gbits/sec. Currently, HIPPI crossbar switches are commercially available in 8 by 8 configurations from Network Systems Corporation (NSC).

Both crossbar switches have full 800 Mbit/sec HIPPI on all ports. The connect time for each is done in a few hundred nanoseconds. The NSC CBS can do a connect in under 200 ns using 15 meter HIPPI cables between both external nodes and the CBS. Disconnect times usually are less than 100 ns but is also dependent on cable length.

The network architecture of HIPPI devices and a CBS is naturally a star network with HIPPI links radiating from the crossbar, or Cross Point (CP), switch. Such a network is know as a CP* net (Fig. 1).

CP* allows multiple HIPPI devices to be networked together but when there are more devices than ports on the switch, another component is needed. If two CBSs are connected directly through a HIPPI port on each, the external nodes have the problem of routing to the other switch if the destination of their data is not on the same CBS they are on. The solution is to place a crossbar interface (CBI) device between the two CBSs. In addition to routing, the CBI can enforce network security. The CBIs can also perform inter-switch

Figure 1: CP* Environment

routing such that many CP* could be connected together in various architectures. Such a network is being called a Multiple Crossbar Network (MCN) (Fig. 2).

Applications

As previously mentioned, imaging and visualization were some of the driving forces in the development of HIPPI. The primary application of imaging over HIPPI will be the playback of animation sequences over a HIPPI framebuffer in the scientist's office. There is a great reluctance to utilize a visualization laboratory: the papers/journals/notebooks etc relating to the science are in their office not the visualization lab, one loses their train of thought when having to walk down the hall (possibly into another building), the spontaneity of quickly reviewing or studying an animation is not possible, etc[4]. Framebuffers directly connected into a particular machine are typically in a very limited number of fixed locations. The capability of having a networked framebuffer overcomes this problem.

Scientists want to view images sequences representing visualizations of their data with a VCR type interface. These image sequences can be generated using post-processing visualization techniques or from graphics which are directly coupled to the model running on a supercomputer (monitoring running models). An obvious option is to stream the frames to the local workstation and play them back locally. This often is unfeasible due to the huge size of the aggregate frames (1 Mbyte/frame) and the small memory size of workstations. A typical interactive session consists of a limited number of small packets controlling the animation (from the scientist to the supercomputer) and a large number of very big packets arriving which contain the frames to be viewed. Data compression can be very useful for the post-processing scenario but for simulation tracking or simulation steering, it poses problems. The time required to perform compression on one side and decompression on the other is

Figure 2: MCN Configuration

the limiting factor. Recently, silicon implementations of compression algorithms have been brought to market (i.e. JPEG). However, these are directed at single frames, not sequences. MPEG addresses image sequences by using both spatial and temporal compression. However when dealing with image compression, one must consider whether the compression technique is losey or loseless. Loseless compression retains the same quantitative information in the decompressed image. Whereas with losey techniques, information content is traded for file size. Both JPEG and MPEG are losey compression techniques. While these are acceptable for video teleconferencing, NTSC images or digital video, with scientific data techniques which modify the quantitative information are unacceptable for many applications. The scientist must not be distracted from examining phenomena by artifacts introduced by a compression/decompression technique. Worst still is the introduction of artifacts which might be misconstrued as phenomena within the data.

We are in the process of developing a general capability where the scientist can review high resolution frames of images (animation) via the HIPPI framebuffer through a VCR type interface. Los Alamos has developed a 1024 by 1024 by 24-bit image HIPPI frame buffer [5]. This 24-bit device can run in two modes: a resolution of 1024 by 1024 at 15 frames/sec and a resolution of 640 by 512 at 60 frames/sec. Currently, this device is driven in the production environment directly off of Cray YMPs. When driven by the YMP, the framebuffer user contends with other concurrent users (timeslicing) as well as I/O subsystem contention. However, we have found the device to give quite acceptable results. It is easy to raise the priority of the framebuffer job to receive a more generous timeslice. However, disk contention is much more difficult schedule. Rather than build a 24-bit movie in memory or on disk, the interface provides for on the fly decoding of 8-bit colormapped image sequences. This reduces

the I/O subsystem requirements. Additionally, we have interfaced, via HIPPI, a RAID disk for caching of animation loops for later and smoother playback at a later time. This provides the best results for postprocessing of data producted by models. For interactive work, model monitoring or simulation steering, the simulation must be paused while the visualization is produced and the image sequence is loaded onto the RAID disk for smooth playback. We have found the attached framebuffer provides better results in this scenario. Since the framebuffer is a component of our HIPPI network, the framebuffer is an addressable network device. Thereby providing HIPPI animation capability into a large number of locations. In addition to the 24-bits of image information per pixel, the framebuffer requires 3 bits of control information, 1 bit of audio, and 4 bits are reserved for future use (multimedia or other). With 1 bit per pixel reserved for audio at 1024x1024, each frame has the capacity of 1Mb of accompanying audio information. At 15 frames/second, this provides 22 channels of CD quality sound (assuming 16 bits/sample and 44,100 samples/second. This currently is not being used in the production environment but research projects are investigating both auditorialization and multimedia[6, 7].

We are also studying distributed visualization via HIPPI. The typical visualization process consists of moving the raw data computed on the supercomputer to a graphics workstation. The data is then culled, filtered, mapped and rendered on the graphics workstation. This can be thought of as a visualization pipeline. In the high-speed network distributed visualization model, parts of the pipeline are migrated to the appropriate hardware within the network. The most obvious is to cull and filter the data on the supercomputer, transport to the graphics workstation where mapping and rendering take place. Although the bandwidth of most workstations' backplanes is lower than HIPPI, HIPPI to VME cards are commercially available. This still remains a bottleneck on the workstation side but can still be very effective with clever partitioning of the visualization pipeline.

Another migration is to perform the mapping on the massively parallel computer and transport geometry via the high-speed network for rendering on the graphics workstation. We have implemented a massively parallel isosurface extraction algorithm, based on Marching Cubes, on the Cm-2 [8]. In this environment, the scientist's model or simulation executes the isosurface extraction algorithm and the resulting geometry is transferred, via a high-speed network, to their workstation, in our case an SGI VGX, for rendering. Figure 3 shows a sequence of rendered images produced with this distributed environment. For a 256x256x256 volume of floating point data, the raw data requires 530Mbits per time step. Considering that dynamic simulations contain hundreds of time-steps, this is obviously too much raw data to transport in the typical visualization process. If 50K polygons (triangles) are extracted, the data shipped over the network is reduced to 14Mbits. This represents of compression factor of almost 37 times! As previously mentioned, due to the VME restrictions on current graphics platforms the network remains the bottleneck for this problem. To help overcome this problem, we have implemented a temporal lossless compression algorithm for transmitting only changed geometry between time-steps. We continue to investigate other mappings of the visualization pipeline onto the high-speed networked environment.

Figure 3: Rendered Polygons

References

[1] Karl-Heintz Winkler. An integrated system approach to large-scale scientific computing. *Los Alamos Colloquium Series*, April 1985.

[2] D. E. Tolmie et. al. Interconnecting computers with the high-speed parallel interface. Technical Report LA-9503-MS, Los Alamos National Laboratory, 1982.

[3] John Fowler and Michael McGowen. Design and implementation of a supercomputer frame buffer system. In *Proceedings of SuperComputing 1988*, pages 140–147, 1988.

[4] Richard L. Phillips. A scientific visualization workbench. In *Proceedings of SuperComputing 1988*, pages 145–148, 1988.

[5] Wally St. John. High-performance parallel interface (hippi) image frame buffer. internal document, Los Alamos National Laboratory.

[6] R. S. Hotchkiss and C. L. Wampler. The auditorialization of scientific information. In *Proceedings of Suprtcomputing Conference 1991*, pages 453–461, 1991.

[7] Richard L. Phillips. Mediaview: A general multimedia digital publication system. *Communications of the ACM*, 34(7), July 1991.

[8] W. Lorensen and H Cline. A high resolution 3d surface contruction algorithm. In *Computer Graphics*, volume 21, pages 163–169, 1987.